Une grille d'évaluation pour les analyseurs syntaxiques

Philippe Blache (1) & Jean-Yves Morin (2)
(1) LPL-CNRS, Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence
pb@lpl.univ-aix.fr
(2) Département de Linguistique, Université de Montréal
CP 6128, succ. Centre-ville
Montréal QC H3C 3J7
morinjy@sympatico.ca

Mots-clefs – Keywords

Evaluation, grille d'analyse, performance, diagnostic
Evaluation template, performance, diagnostic

Résumé - Abstract

Les techniques d'évaluation aujourd'hui disponibles posent un certain nombre de problèmes à la fois pour ce qui concerne la disponibilité des ressources nécessaires, mais également dans la mesure où elles ne reflètent pas complètement des véritables capacités d'un système. Nous proposons dans cet article l'élaboration d'une grille d'analyse qui constitue une description standardisée des caractéristiques du système. Cette grille contient à la fois des aspects purement descriptifs (concernant par exemple le formalisme ou les aspects algorithmiques) ainsi qu'un ensemble de mesures automatisées, le tout donnant une image plus précise du système qu'une simple évaluation quantitative.

The evaluation procedures that have been tested especially for english or german pose some problems. First, resources required for a quantitative evaluation (typically a treebank) are not well developped for french. Moreover, such measures only concern subset of the system capabilities. We propose in this paper an evaluation template containing both descriptive characteristics (presentation of the formalism, faithfulness of the implementation, algorithmic aspects) and automatic measurements (recall, precision, relations, tests suites, etc.).

1 Introduction

Les résultats des précédentes campagnes d'évaluation, en particulier pour ce qui concerne les méthodologies mises en œuvre, doivent être analysée de près avant de s'engager dans le développement d'un nouveau protocole d'évaluation en France. De ce point de vue, il est intéressant de revenir d'une part sur les campagnes engagées par l'AUPELF dans le milieu des années 90 (cf. (8) dont les résultats ont joué un rôle majeur dans la création de la conférence LREC, ainsi bien entendu que sur les campagnes d'évaluations proposées pour l'anglais et en particulier sur les résultats de Parseval (cf. (12)).

Sans entrer dans le détail de tous les projets, il convient, à titre de remarque préliminaire, de constater qu'un même modèle d'évaluation ne peut convenir à n'importe quel système de traitement de la langue. Les systèmes de traitement de la parole étaient par exemple dans les années 90 bien plus avancés que les systèmes de traitement de l'écrit et pouvaient se prêter plus facilement à des évluation quantitatives. Pour ce qui concerne l'écrit, s'il était possible d'évaluer et de comparer de façon assez précise des systèmes d'étiquetage morpho-syntaxiques (cf. campagne GRACE), il était en revanche beaucoup plus difficile (pour ne pas dire impossible) de dresser ne serait-ce qu'un véritable protocole pour l'évaluation des systèmes de compréhension (cf. (2)). L'évaluation des analyseurs syntaxiques n'échappe pas à cette règle et les activités menées dans le cadre de Parseval ont montré les mêmes limites. Il n'est pas possible de réduire l'évaluation de tels systèmes à une simple évaluation quantitative reposant sur la détection de frontières et de types d'unités. Encore faut-il également savoir de quelles unités parle-t-on ainsi que du niveau d'information nécessaire à l'analyse. Ainsi, un analyseur s'appuyant exclusivement sur les informations syntaxiques sans autre type d'information (par exemple la sémantique lexicale), devra donner comme résultat l'ensemble des structures syntaxiques possibles. Il s'agit donc d'un premier problème ayant des conséquences importantes sur les ressources utilisées pour l'évaluation : un corpus de référence annoté sur la base duquel sont évalués les résultats devra par exemple comporter cet ensemble d'analyses. Le problème se complique encore si on cherche à comparer des analyseurs. Il existe en effet un grand nombre de formalismes linguistiques et il n'est pas possible de concevoir une représentation générique de l'information syntaxique qui serait utilisable quel que soit le formalisme. Enfin, et cette question relève tout autant du génie logiciel, il est indispensable de connaître le niveau de généralité du système, ainsi que ses capacités d'évolution et d'adaptation.

Nous allons dans cet article revenir sur l'ensemble de ces points. Il est clair qu'une campagne d'évaluation doit prendre en compte un certain nombre de critères dépassant largement la stricte évaluation quantitative. Mais il est dans le même temps indispensable de s'interroger non seulement sur les ressources nécessaires à une telle évaluation ainsi que sur notre capacité à les développer. Plus généralement, encore, la question préliminaire, avant de savoir comment évaluer des systèmes, est de savoir pourquoi évaluer ces systèmes et qu'entend-on par évaluation. La réponse à ces questions permet de mettre en perspective l'utilisation de techniques et ressources existantes tout en les situant dans un cadre plus général. En d'autres termes, nous proposons une approche permettant de réutiliser les protocoles existants en offrant une possibilité d'interprétation dans un cadre générique.

2 Situation

Nous nous proposons dans cette section de faire le point sur les différentes techniques ou proposition faites dans le domaine de l'évaluation.

Les critères d'évaluation des analyseurs syntaxique et plus généralement des systèmes de traitement des langues naturelles (cf. notamment (**?**), (20) ou (2)) sont de plusieurs types. Ils portent bien entendu sur des mesures quantitatives, mais également sur l'adéquation du système aux objectifs visés ainsi qu'à sa capacité d'évolution. Nous décrivons plus particulièrement dans cette section les principales approches portant sur le comportement et les résultats fournis par un système.

Les critères classiques de l'évaluation d'applications, typiquement le rappel et la précision, permettent de situer les résultats par rapport à un standard. La quantification de ces critères se fait donc sur la base d'un corpus annoté, typiquement un corpus arboré du type Penn treebank (cf. (18)) ou Susanne (cf. (19)). On parle dans ce cas d'évaluation de la performance d'un système. On utilise pour cela des procédures entièrement automatisées.

Mais il existe également un second type d'évaluation qui relève plutôt du diagnostic. Il s'agit dans ce cas d'examiner les résultats fournis par un système pour l'analyse d'un ensemble de phrases tests comportant des difficultés ou des tournures particulières, y compris non grammaticales.

2.1 Evaluation sur corpus arborés

Ce type de corpus arboré propose essentiellement l'annotation d'informations portant sur le type des catégories, leur étendue ainsi que d'autres informations complémentaires (mais de façon moins systématique) concernant par exemple la structure prédicative. Il s'agit donc typiquement de parenthésage encodant essentiellement des informations sur la constituance ainsi que quelques relations syntaxiques à proprement parler. Les structures proposées dans ce type de corpus sont plates plutôt que profondes, et le nombre d'emboîtements est limité. Ce choix est essentiellement fonctionnel et favorise généralement le consensus entre annotateurs. Cette caractéristique implique l'utilisation de catégories de niveau relativement haut de préférence à des catégories de granularité plus fine (cf. Blache98). Par ailleurs, les informations encodées relèvent exclusivement du niveau syntaxique. Enfin, une structure syntaxique est choisie pour être encodée dans le cas d'analyses multiples, au détriment de la représentation de l'ambiguïté. Ce premier type de corpus, de loin le plus répandu pour l'anglais, repose donc sur un certain nombre de choix concernant à la fois le formalisme linguistique ainsi que le type d'information à encoder.

La première technique d'évaluation consiste simplement à mesurer dans quelle mesure un analyseur est capable de reproduire les informations contenues dans ce type de corpus. C'est typiquement le cas des évaluations menées dans le cadre de Parseval qui propose de mesurer la précision, le rappel ainsi que le nombre de croisements de parenthèses entre le corpus de référence et la sortie de l'analyseur.

Plusieurs critiques sont faites à ce type d'évaluation. Tout d'abord, et c'est la critique principale, les mesures proposées sont valides pour des analyseurs syntagmatiques. Un certain nombre d'annotations notamment du Penn treebank relèvent encore plus précisément de choix

théoriques et formels pertinents dans un cadre particulier, mais peu adaptés pour d'autres approches. Par ailleurs le choix des étiquettes associées aux catégories, en d'autre termes le typage des objets manipulés, est d'une granularité élevée, ce qui ne permet pas toujours la description de phénomènes variés. Plusieurs auteurs (cf. par exemple (**?**)) ont montré que ce type de mesures pénalise les analyseurs proposant des annotations plus précises (donc un plus grand nombre de parenthèses que dans le corpus de référence).

2.2 Evaluation sur des suites de phrases

Plusieurs projets ou campagnes d'évaluation ont proposé de tester les performances des systèmes sur la base d'un ensemble de phrases tests. La plupart des grands projets de développement d'analyseurs syntaxiques ont élaboré leur propre jeu de phrases. C'est le cas par exemple du projet Alvey (cf. (5)) proposant des analyseurs basés sur GPSG ou d'un projet similaire mené par le laboratoire Hewlett-Packard (cf. (10)). On doit également signaler en France une première opération de comparaison de plusieurs analyseurs pour le français proposant également un tel jeu (cf. (9)).

Mais l'opération la plus systématique en termes de ressources développées, de couverture et de nombre de langue reste bien entendu TSNLP (cf. (14) et (15)). Ce projet a proposé une méthodologie, des conventions d'annotation ainsi que des outils pour l'élaboration de tels jeux. L'idée est d'identifier un certain nombre de phénomènes linguistiques et pour chacun d'entre eux, de proposer une série de phrases. Une des particularités de TSNLP est de proposer également des phrases mal formées, ce qui permet de tester le comportement de l'analyseur en terme de reconnaissance mais également de robustesse.

Parmi les phénomènes listés, notamment pour le français, nous trouvons la complémentation, la modification, l'accord, la coordination, la négation, etc. Ce type d'information est utile pour une évaluation dite *diagnostique* du système mais également indispensable pour analyser l'évolution du développement de la grammaire et de l'analyseur.

2.3 Evaluations relationnelles

Un des problèmes majeurs posés à l'évaluation vient du fait que, à la différence des étiqueteurs morphologiques, l'analyse syntaxique repose sur des formalismes variés. Il est donc impossible de prétendre représenter des structures syntaxiques complètes sans être dépendant d'un formalisme donné. Ainsi, la plupart des corpus annotés s'appuient sur des grammaires syntagmatiques. Il n'est donc pas possible d'utiliser *directement* ces corpus pour évaluer des analyseurs utilisant d'autres formalismes comme les grammaires de dépendance par exemple.

Dans la mesure où il n'est pas concevable ni d'un point de vue matériel ni même d'un point de vue d'efficacité de développer des ressources adaptées à des formalismes, il convient de chercher une démarche intermédiaire. Une solution proposée consiste à exploiter pour l'évaluation non pas les structures syntaxiques (en particulier les unités et leurs frontières), mais plutôt les relations syntaxiques. On trouve des propositions allant dans ce sens dans (13) ou encore Lin02 qui suggère d'utiliser des relations de dépendances plutôt que de constituance y compris pour évaluer des analyseurs syntagmatiques. Dans cette perspective, chaque mot est associé à trois types d'informations :

- la catégorie du mot courant

- la tête modifiée par le mot courant et sa localisation par rapport à celui-ci

- le type de relation qui unit les deux mots : sujet, adjoint, complément, spécifieur, etc.

Une autre approche plus systématique et encore plus générale propose de s'appuyer sur les relations grammaticales plutôt que sur d'autres types d'informations. Il s'agit du schéma d'annotation de relations grammaticales (cf. (6), (7)). Les auteurs proposent l'annotation d'un ensemble de relations syntaxiques identifiables indépendamment du formalisme choisi. De plus, ces relations sont hiérarchisées ce qui permet des niveaux de précision différents dans l'annotation. Nous rappelons brièvement le type de relations proposées par ces auteurs.

Niveau	Nom	Arguments	Description
1	*dependent*	introducteur tête dépendant	Relation de dépendance générique entre une tête et un dépendant
1.1	*mod*	type tête dépendant	Relation entre une tête et son modifieur. Le type est le mot introduisant la dépendance
1.1.1	*ncmod*	-	Modificateur lexical (non clausal)
1.1.2	*xmod, cmod*	-	Modificateurs propositionnels
1.2	*arg-mod*	type tête dépendant rel-initiale	Relation tête/argument, celui-ci étant réalisé comme un modifieur
1.3	*arg*	tête dépendant	Relation générique tête/argument (plutôt de type complément)
1.3.1	*subj*	tête dépendant relation	Relation prédicat/sujet
1.3.1.1	*ncsubj*	-	Sujet lexical (non clausal)
1.3.1.2	*xsubj, csubj*	-	Sujets propositionnels
1.3.2	*comp*	tête dépendant	Relation tête/complément
1.3.2.1	*obj*	-	Relation tête/objet
1.3.2.1.1	*dobj*	-	Relation prédicat/objet direct (premier complément non propositionnel)
1.3.2.1.2	*iobj*	-	Relation prédicat/complément non propositionnel introduit par une préposition
1.3.2.1.3	*obj2*	-	Relation prédicat/second complément non propositionnel
1.3.2.2	*clausal*	-	Relation tête/complément propositionnel

Il est important de rappeler pour terminer, si besoin était, que le type de ressource que nous venons de décrire n'existe que marginalement pour le français. Une seule véritable ressource a été développée sous l'impulsion de Anne Abeillé qui a, avec son équipe, produit un corpus arboré sur la base de textes issus du journal "Le Monde". L'exemple suivant donne un extrait de cette ressource pour la phrase "*Seuls pirates, marchands d'esclaves et trafiquants de drogue y sont légitimement poursuivis par tous*" (les balises fermantes sont omises pour des raisons de lisibilité).

```
<SENT nb="10000">
   <NP>
      <w lemma="seul" ei="AImp" ee="A-ind-mp" cat="A" subcat="ind" mph="mp">Seuls</w>
      <w lemma="pirate" ei="NCmp" ee="N-C-mp" cat="N" subcat="C" mph="mp">pirates</w>
   <COORD>
      <w lemma="," ei="PONCTW" ee="PONCT-W" cat="PONCT" subcat="W">,</w>
   <NP>
      <w lemma="marchand" ei="NCmp" ee="N-C-mp" cat="N" subcat="C" mph="mp">marchands</w>
      <PP> <w lemma="de" ei="P" ee="P" cat="P">d'</w>
         <NP> <w lemma="esclave" ei="NCmp" ee="N-C-mp" cat="N" subcat="C" mph="mp">esclaves</w>
...
   <VN>
      <w lemma="y" ei="CL3fs" ee="CL–3fs" cat="CL" subcat="" mph="3fs">y</w>
      <w lemma="être" ei="VP3p" ee="V–P3p" cat="V" subcat="" mph="P3p">sont</w>
      <w lemma="légitimement" ei="ADV" ee="ADV" cat="ADV">légitimement</w>
      <w lemma="poursuivre" ei="VKmp" ee="V–Kmp" cat="V" subcat="" mph="Kmp">poursuivis</w>
   <PP> <w lemma="par" ei="P" ee="P" cat="P">par</w>
      <NP> <w lemma="tous" ei="PROmp" ee="PRO–3mp" cat="PRO" subcat="" mph="3mp">tous</w>
</SENT>
```

2.4 Quelques informations de base pour la grille

2.5 Evaluation des aspects syntagmatique

L'analyse de la situation décrite précédemment nous permet de dégager quelques directions de réflexion. Les expériences montrent tout d'abord qu'il ne faut pas s'engager dans l'annotation d'informations syntaxiques en utilisant un formalisme trop spécifique pas plus qu'on ne peut prétendre à proposer une annotation générique. Mais plusieurs approches ont montré qu'il était possible d'extraire un certain type d'information, plutôt de niveau relationnel, à partir d'encodages variés.

Par ailleurs, les travaux récents ont montré les limites d'une seule évaluation des performances sur la base d'un parenthésage. Ce type d'information, en plus du fait qu'il est dépendant d'un formalisme, pose de nombreux problèmes, notamment pour une évaluation précise. Mais là encore, il ne faut pas négliger le fait que la plupart des analyseurs superficiels (donc une bonne proportion des systèmes d'analyse aujourd'hui disponibles) produisent de telles structures. Il est donc nécessaire de préserver la possibilité d'exploiter ces informations. Il faut cependant compléter ces aspects par une granularité plus fine des mesures. Imaginons par exemple que nous ayons 85% de SN de la forme Dét+N. Le parenthésage et l'étiquetage incorrect des SN qui ne sont pas de ce type devrait donc tenir compte de cet aspect et l'évaluation devrait être pondérée en fonction de la fréquence et/ou de la complexité de la structure.

A ce stade, il est nécessaire d'aborder le problème de la distinction analyseur/grammaire. En effet, tous les protocoles évaluent aujourd'hui conjointement ces deux composants. Mais il est intéressant et sans doute important d'entrer dans un niveau de description plus fin, y compris en termes computationnels. Les analyseurs symboliques distinguent les parties données et traitement. Une grammaire est dans ce cas clairement distincte de l'analyseur qui l'utilise. Même si une telle distinction n'est pas valide pour d'autres techniques, il faut pouvoir distinguer dans l'évaluation la qualité de la grammaire de celle de l'analyseur à proprement parler. Il s'agit bien entendu d'une entreprise extrêmement difficile que de tenter de caractériser ces deux aspects, souvent indissociables. On peut toutefois citer un certain nombre de carac-

téristiques propres à l'analyseur. Il est important par exemple de juger du comportement du système pour le traitement des non attendus, sa robustesse. Il est également important de prendre en compte le déterminisme de l'analyse et la capacité à hiérarchiser les solutions en cas de non-déterminisme. Sans entrer dans une tentative de quantification de ces aspects, un élément d'information intéressant réside dans le type et la quantité de structures intermédiaires produites en cours d'analyse (par exemple, pour un analyseur tabulaire, le nombre total d'arcs utilisés pour une analyse donnée). Pour ce qui concerne la grammaire, l'évaluation distincte est difficile à cerner en dehors d'une comparaison entre plusieurs versions d'une grammaire. D'un point de vue purement descriptif, il est malgré tout important de connaître le nombre et le type de catégories utilisées, le niveau de hiérarchisation (s'il existe) ou le type d'encodage de l'information (règles, contraintes, etc.).

Le dernier aspect qui nous semble déterminant à prendre en compte concerne le formalisme lui-même. Il est en effet, nous y reviendrons dans la section suivante, important de connaître le paradigme théorique dans lequel se situe l'analyseur et surtout, puisqu'un des aspects de l'évaluation est la comparaison entre plusieurs systèmes, de connaître le degré de fidélité de l'implantation à cette théorie. Nous sommes ainsi en mesure de savoir, dans le cas où une théorie aurait un intérêt particulier au-delà de la syntaxe par exemple, s'il est possible d'exploiter la totalité de son pouvoir expressif ou pas.

2.6 Informations non syntagmatiques

Les informations purement syntagmatiques sont bien entendu fondamentales, mais il est également très important d'évaluer les autres dimensions comme les fonctions grammaticales, les rôles abstraits, la structure communicative, la structure anaphorique, la deixis, etc. Là encore, il ne s'agit pas de limiter l'évaluation aux simples procédures de mesures quantitatives. Même si nous ne disposons pas de corpus de référence encodant ce type d'information, il est important de disposer d'une description précise de la prise en compte par le système de ces informations.

Par ailleurs, la syntaxe entretient des liens étroits avec d'autres domaines linguistiques. Il est donc nécessaire de prendre en compte la capacité de l'approche et du système en particulier à représenter et traiter les relations avec d'autres domaines proches comme la morphologie ou la sémantique. Cette question d'interfaçage est déterminante pour plusieurs raisons. D'une part, elle fournit une indication sur les potentialités d'utilisation du système. D'autre part, un analyseur ouvert sur d'autres domaines est révélateur d'une technologie plus durable et à terme plus efficace. A un niveau plus général, ces paramètres nous semblent concourir à une évaluation, ou du moins une spécification, des capacités d'évolution du système.

3 Une grille d'évaluation des analyseurs

Avant de décrire plus précisément la grille d'évaluation que nous proposons, il est important de tirer quelques conclusions des remarques faites précédemment. Tout d'abord, face à la difficulté de la tâche de constitution de ressources pour l'évaluation (en particulier les corpus annotés), il semble nécessaire de tirer parti des toutes les ressources disponibles, quelles qu'elles soient. Par ailleurs, il semble tout aussi nécessaire de ne pas limiter l'évaluation aux simples mesures de performance de l'analyseur. Il convient de situer cette remarque dans la perspective d'une question plus générale : à quoi sert l'évaluation de tels systèmes ? La réponse varie en fonction

de la destination : elle sert de validation pour le développeur lui-même, mais elle est aussi indicatrice des capacités du système pour un utilisateur potentiel qui chercherait le système le plus adapté à ses besoins. Il faut donc compléter les mesures ou benchmarks par une description du système à proprement parler. Nous proposons donc de distinguer deux types d'informations dans la grille d'évaluation :

- les évaluations quantitatives : ensemble d'opérations automatiques ou semi automatiques mesurant des résultats d'analyse sur des corpus de référence annotés ou des ensembles de phrases-tests. Une description analytique des sorties pourra compléter les mesures (typiquement le comportement du système sur des entrées mal formées),

- la description du système : informations fournies par le concepteur du système concernant les aspects non mesurables automatiquement.

On récapitule dans le tableau suivant l'ensemble des caractéristiques évoquées précédemment et qui nous semblent devoir entrer en ligne de compte pour une évaluation précise, voire une comparaison des systèmes d'analyse syntaxique. Il ne s'agit pas bien entendu d'une liste exhaustive. De même en fonction des objectifs de l'évaluation, il peut ne pas être utilse de renseigner tous les champs de la grille. Il nous semble cependant utile de founir une vision aussi précise du système qui permette de mettre en perspective les mesures quantitatives par rapport à la base théorique et computationnelle du système.

Description	• Formalisme, théorie – Description du cadre théorique – Description de la fidélité de l'implantation à ce cadre théorique, analyse des simplifications si elles existent. • Description du type d'information retourné par le système – Catégories : types de catégorie, granularité, représentation – Représentation de l'information syntaxique : règles, relations, etc. – Structures construites : arbres, graphes, ensembles, etc. • Ressources utilisées : – Lexiques, grammaires – Outils • Description algorithmique – Architecture – Stratégie d'analyse, déterminisme – Hiérarchisation des solutions – Complexité

Le tableau présenté ci-après récapitule l'ensemble des mesures qu'il nous semble intéressant de faire. Là encore, il ne s'agit pas d'une liste exhaustive prétendant évaluer précisément tous les aspects du système. Il nous semble cependant important, compte tenu de toutes les critiques faites aux autres protocoles, de proposer une approche mixte, tirant parti de toutes les techniques existantes dans ce domaine. Une telle démarche permet de plus de tirer le meilleur parti des ressources existantes sans les mettre en concurrence.

Mesures	• Parenthésage – Description du corpus de référence – Rappel – Précision – Croisements • Relations – Définition del'ensemble des relations à évaluer – Description du corpus de référence – Quantification • Phrases tests – Définition de l'ensemble des tournures visées – Description du jeu de phrases – Quantification – Description des analyses pour les entrées mal formées

La partie d'évaluation des relations mérite quelques commentaires. Il s'agit là de s'inscrire dans la démarche proposée par (7) et donc de spécifier une ensemble de relations syntaxiques nous semblant refléter à la fois de la complexité du problème ainsi que des capacités au moins partielle des systèmes. La liste des relations n'est pas figée. La proposition récapitulée dans la section précédente constitue une base de départ qui peut éventuellement être complétée voire adaptée pour le français (certaines relation peuvent en effet être pertinentes essentiellement pour l'anglais).

4 Conclusion

Les techniques d'évaluation aujourd'hui disponibles posent un certain nombre de problèmes. Tout d'abord, une évaluation purement quantitative ne permet pas de caractériser précisément les capacités d'un système. De plus, la disponibilité des ressources nécessaires pour une telle évaluation est contrastée selon les langues. Avant de s'engager dans une campagne d'évaluation plus systématique des analyseurs existants pour le français, il nous a donc semblé utile de faire le point de la situation et de proposer, plutôt qu'une métrique ou u pvéritable protocole, un cadre général que nous appelons grille d'évaluation. L'idée est d'une part de rendre compte d'aspects généraux caractérisant l'analyseur (par exemple le formalisme choisi ou encore certaines caractéristiques algorithmiques) et d'autre part de rassembler plusieurs techniques d'évaluation (diagnostic, performances, etc.). Une telle approche nous semble être raisonnable dans la mesure où, plutôt que d'élaborer une théorie de l'évaluation, nous rassemblons plusieurs indices ou critères caractérisant au mieux les systèmes.

Références

1. [Atwell96] Atwell E. (1996) "Comparative evaluation of grammatical annotation models" In R. Sutcliffe, H. Koch, A. McElligott (Eds.), *Industrial Parsing of Software Manuals*, Rodopi.

2. [Blache97] Blache P., J. Guizol, F. Lévy, A. Nazarenko, S. N'Guema, M. Rolbert, R. Pasero & P. Sabatier (1997)"Evaluer des systèmes de compréhension de textes ", in Actes des Journées Scientifiques et Techniques (JST 97, Avignon), AUPELF-UREF.

3. [Briscoe95] Briscoe, E., Carroll, J. (1995) “Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels”, in Proceedings of *the 4th International Workshop on Parsing Technologies*.

4. [Carpenter97] Carpenter B. & Manning, C. (1997) “Probabilistic parsing using left corner language models”, in Proceedings of the *5th ACL/SIGPARSE International Workshop on Parsing Technologies*.

5. [Carroll91] Carroll, J., E. Briscoe & C. Grover (1991) “A development environment for large natural language grammars”, Computer Laboratory, Cambridge University, UK, Technical Report 233.

6. [Carroll98] Carroll, J., Briscoe E. & Sanfilippo, A. (1998) “Parser evaluation: a survey and a new proposal”, in Proceedings of the *International Conference on Language Resources and Evaluation*.

7. [Carroll02] Carroll J., G. Minnen & T. Briscoe (2002) “Parser evaluation: using a grammatical relation annotation scheme”, in A. Abeillé (ed), *Trebanks: Building and Using Syntactically Annotated Corpora*, Kluwer.

8. [Chibout00] Chibout K., F. Néel, J. Mariani et N. Masson (eds) (2000) *Ressources et Evaluation en Ingénierie de la Langue*. Duculot / De Boeck-Université.

9. [Fay-Varnier91] Fay-Varnier C., C. Fouqueré, G. Prigent, & P. Zweigenbaum (1991) “Modules syntaxiques des systèmes d’analyse du français”, in *Technique et Science Informatiques*, 10(6).

10. [Flickinger87] Flickinger D., J. Nerbonne, I. Sag & T. Wasow (1987) “Toward Evaluation of NLP Systems”, HP Labs Technical Report, Palo Alto.

11. [Gaizauskas98] Gaizauskas, R., Hepple M., Huyck, C. (1998) “Modifying existing annotated corpora for general comparative evaluation of parsing”, in Proceedings of the *LRE Workshop on Evaluation of Parsing Systems*.

12. [Harrison91] Harrison, P., Abney, S., Black, E., Flickinger, D., Gdaniec, C., Grishman, R., Hindle, D., Ingria, B., Marcus, M., Santorini, B., Strzalkowski, T. (1991) “Evaluating syntax performance of parser/grammars of English”, in Proceedings of the *Workshop on Evaluating Natural Language Processing Systems*, 29th Annual Meeting of the Association for Computational Linguistics.

13. [Kübler02] Kübler S. & H. Telljohann (2002) “Towards a Dependency-Oriented Evaluation”, in Proceedings of *Beyond Parseval Workshop*.

14. [Lehmann96a] Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., Arnold, D. (1996) “TSNLP - test suites for natural language processing”, in Proceedings of *COLING’96*.

15. [Lehmann96b] Lehmann S., D. Estival & S. Oepen (1996) “TSNLP - test suites for natural language processing”, in Proceedings of *COLING’96*.

16. [Lin98] Lin, D. (1998) “A dependency-based method for evaluating broad-coverage parsers”, in *Natural Language Engineering*.

17. [Lin02] Lin, D. (2002) “Dependency-based evaluation of MINIPAR”, in A. Abeillé (ed), *Trebanks: Building and Using Syntactically Annotated Corpora*, Kluwer.

18. [Marcus93] Marcus, M., Santorini, B., Marcinkiewicz, M. (1993) “Building a large annotated corpus of English: The Penn Treebank”, in *Computational Linguistics*, 19(2).

19. [Sampson95] Sampson G. (1995) *English for the Computer: The SUSANNE Corpus and Analytic Scheme* Oxford University Press.

20. [Sparck Jones96] Sparck Jones K. & J. Galliers (1996) *Evaluating Natural Language Processing Systems*, Springer.

21. [Srinivas96] Srinivas, B., Doran, C., Hockey B., Joshi A. (1996) “An approach to robust partial parsing and evaluation metrics”, in Proceedings of *the ESSLLI’96 Workshop on Robust Parsing*.